# On-Line Learning with Time-Correlated Patterns

Wim Wiegerinck and Tom Heskes

Department of Medical Physics and Biophysics, University of Nijmegen,
Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands.

**Abstract**

Current theories on on-line learning in neural networks are based on the unrealistic assumption that subsequent patterns are uncorrelated. In this paper we study on-line learning with time-correlated patterns. For small learning parameters we derive a Fokker-Planck equation describing the evolution of the average network state and the fluctuations around this average. Correlations between subsequent patterns contribute to the diffusion term in this Fokker-Planck equation and thus affect the fluctuations in the learning process. Our results are valid for a general class of learning rules, including backpropagation and the Kohonen learning rule. Simulations with Oja's rule illustrate the theoretical results.

PACS: 87.10., 02.50., 05.40.

*Europhys. Lett.*, **28** (6), pp. 451-455 (1994)

On-line learning is a learning process for neural networks where the weights of the network are updated each time a training pattern $\vec{x}$ is presented to the network. For many learning processes this weight change can be written in the general form

$$\Delta \boldsymbol{w}(n) \equiv \boldsymbol{w}(n+1) - \boldsymbol{w}(n) = \eta \, \boldsymbol{f}(\boldsymbol{w}(n), \vec{x}) \,, \tag{1}$$

with $\boldsymbol{w}(n)$ the network state at iteration step $n$, $\eta$ the learning parameter, and $\boldsymbol{f}(\cdot,\cdot)$ the learning rule. Examples can be found in supervised learning, *e.g.*, backpropagation for multilayer perceptrons [1], where $\vec{x}$ stands for an input and desired output pair, as well as in unsupervised learning, *e.g.*, Kohonen's self-organising rule for topological feature maps [2], where $\vec{x}$ stands for the input vector.

If the patterns $\vec{x}$ are drawn at random from the training set, on-line learning (1) can be described by a first-order Markov process, since the new network state $\boldsymbol{w}(n+1)$ is solely a function of the old state $\boldsymbol{w}(n)$ and the randomly drawn pattern $\vec{x}$. An evolution equation for the probability $P(\boldsymbol{w}, n)$ that at step $n$ the network is in state $\boldsymbol{w}$ follows directly from (1). In recent years, theoretical studies of provided a better understanding of on-line learning processes with uncorrelated patterns [3 – 7].

However, in biological learning as well as in real world applications subsequent patterns are correlated. Clearly, the existing theory based on random pattern presentation is not valid in such cases. In this paper, we therefore study on-line learning (1) with time-correlated patterns, and we show that for small learning parameters the behaviour of $P(\boldsymbol{w}, n)$ can still be analysed. For the moment, we assume that the probability that a pattern $\vec{x}$ is presented to the network depends on its predecessor $\vec{x}'$ through a transition probability $\rho(\vec{x}|\vec{x}')$, *i.e.*, that the patterns follow a first-order Markov process. Later we will show that our final results hold for stationary Markov processes of any finite order.

With time-correlated patterns the dynamics in weight space is no longer Markovian, which makes it much less straightforward to derive an evolution equation for $P(\boldsymbol{w}, n)$. However, the joint probability $\hat{P}(\boldsymbol{w}, \vec{x}, n)$ that at step $n$ the network is in state $\boldsymbol{w}$ and the presented pattern is $\vec{x}$ *does* follow a Markov process:

$$\Delta \hat{P}(\boldsymbol{w}, \vec{x}, n) = \int d\boldsymbol{w}' \, d\vec{x}' \, \rho(\vec{x}|\vec{x}') \, \delta(\boldsymbol{w} - \boldsymbol{w}' - \eta \boldsymbol{f}(\boldsymbol{w}', \vec{x}')) \, \hat{P}(\boldsymbol{w}', \vec{x}', n) - \hat{P}(\boldsymbol{w}, \vec{x}, n) \,. \tag{2}$$

The time scale for the dynamics of the weights $\boldsymbol{w}$ is inversely proportional to $\eta$, whereas the time scale for the dynamics of the patterns $\vec{x}$ is completely independent of $\eta$. For $\eta \to 0$ this separation of time scales makes it possible to eliminate the fast variable $\vec{x}$ [8] and to derive a systematic expansion of the evolution equation for $P(\boldsymbol{w}, n) = \int d\vec{x} \, \hat{P}(\boldsymbol{w}, \vec{x}, n)$. In [9] we used a similar method to analyse learning with momentum.

To write (2) in a form which is more convenient for algebraic manipulations, we introduce the left and right eigenfunctions $\Psi_i(\vec{x})$ and $\Phi_i(\vec{x})$ of the transition probability $\rho(\vec{x}|\vec{x}')$ with corresponding eigenvalues $\lambda_i$:

$$\int d\vec{x} \, \Psi_i(\vec{x}) \, \rho(\vec{x}|\vec{x}') = \lambda_i \, \Psi_i(\vec{x}') \,, \quad \int d\vec{x}' \, \rho(\vec{x}|\vec{x}') \, \Phi_i(\vec{x}') = \lambda_i \, \Phi_i(\vec{x}) \quad \text{and} \quad \int d\vec{x} \, \Psi_i(\vec{x}) \, \Phi_j(\vec{x}) = \delta_{ij} \quad \forall_{i,j \geq 0} \,.$$

For convenience, we assume that the stationary distribution $\Phi_0(\vec{x})$ in pattern space is unique, *i.e.*, we have $\lambda_0 = 1$, Re $\lambda_{\rm i} < 1$ for $i \geq 1$, and normalisation $\Psi_0(\vec{x}) = 1$. The probability distribution $\hat{P}(\boldsymbol{w}, \vec{x}, n)$ can be decomposed in right eigenfunctions:

$$\hat{P}(\boldsymbol{w}, \vec{x}, n) = \sum_{i \geq 0} Q_i(\boldsymbol{w}, n) \, \Phi_i(\vec{x}) \,.$$

Substitution of this decomposition into (2), multiplication with $\Psi_i(\vec{x})$, and integration over $\vec{x}$ yields a set of equations for $Q_i(\boldsymbol{w}, n)$ [which are treated as components of the vector $\vec{Q}(\boldsymbol{w}, n)$]. Performing a Kramers-Moyal expansion [10] with respect to $\boldsymbol{w}$, this set of equations can be written

$$\Delta \vec{Q}(\boldsymbol{w}, n) = \left[\mathcal{H}\vec{Q}\right](\boldsymbol{w}, n) = \sum_{\alpha=0}^{\infty} \eta^{\alpha} \left[\mathcal{H}^{(\alpha)}\vec{Q}\right](\boldsymbol{w}, n) \,, \tag{3}$$

in which the components of the matrices $\mathcal{H}^{(\alpha)}$ are defined by the differential operators

$$\begin{aligned}\mathcal{H}^{(0)}_{ij} &\equiv -(1-\lambda_i)\delta_{ij} \ , \\ \mathcal{H}^{(\alpha)}_{ij} &\equiv \frac{(-1)^\alpha}{\alpha!}\int d\vec{x}\ \lambda_i \Psi_i(\vec{x})\Phi_j(\vec{x}) \sum_{i_1,\ldots,i_\alpha} \frac{\partial}{\partial w_{i_1}}\cdots\frac{\partial}{\partial w_{i_\alpha}} f_{i_1}(\boldsymbol{w},\vec{x})\cdots f_{i_\alpha}(\boldsymbol{w},\vec{x}), \quad \text{for } \alpha \geq 1.\end{aligned}$$

Since the operator $\mathcal{H}$ is a series in the small parameter $\eta$, it is possible to analyse (3) using perturbation theory. Let us first consider the unperturbed ($\eta = 0$) system

$$\Delta\,\vec{Q}(\boldsymbol{w},n) = \left[H^{(0)}\vec{Q}\right](\boldsymbol{w},n) \ \text{ with solution } \ Q_i(\boldsymbol{w},n) = \lambda_i^n Q_i(\boldsymbol{w},0)\,.$$

The components $Q_i$ with $i \geq 1$ will rapidly relax to zero. For large $n$, only the component $Q_0$ will remain. The eigenvalues of the perturbed system are equal to the eigenvalues $-(1-\lambda_i)$ of the unperturbed system plus terms of order $\eta$. So, if $\eta \ll \min_{i\geq 1}(1 - \mathrm{Re}\ \lambda_i)$, we can still distinguish the invariant subspace in which $\vec{Q}$ rapidly relaxes to zero from the invariant subspace in which $\vec{Q}$ slowly evolves. To describe the evolution of $Q^{\mathrm{s}}$, defined as the projection of $\vec{Q}$ on the slow subspace, we can immediately use a standard result from second-order perturbation theory with degenerate eigenvalues [11],

$$\Delta Q^{\mathrm{s}}(\boldsymbol{w},n) \ = \ \left[\eta\mathcal{H}^{(1)}_{00} + \eta^2\mathcal{H}^{(2)}_{00} + \eta^2\sum_{j>0}\frac{\mathcal{H}^{(1)}_{0j}\mathcal{H}^{(1)}_{j0}}{1-\lambda_j} + \mathcal{O}(\eta^3)\right] Q^{\mathrm{s}}(\boldsymbol{w},n)\,.$$

For large $n$, when the projection of $\vec{Q}$ on the fast subspace has vanished, $Q^{\mathrm{s}}$ equals $Q_0 + \mathcal{O}(\eta^2)$. Substitution of $\mathcal{H}^{(1)}$, $\mathcal{H}^{(2)}$, and $Q_0(\boldsymbol{w},n) = P(\boldsymbol{w},n)$ then leads to

$$\begin{aligned}\Delta P(\boldsymbol{w},n) \ = \ &-\eta\,\boldsymbol{\nabla}_w^T\int d\vec{x}\,\Phi_0(\vec{x})\,\boldsymbol{f}(\boldsymbol{w},\vec{x})\,P(\boldsymbol{w},n) \ + \ \frac{\eta^2}{2}\,\mathrm{Tr}\,\boldsymbol{\nabla}_w\boldsymbol{\nabla}_w^T\int d\vec{x}\,\Phi_0(\vec{x})\,\boldsymbol{f}(\boldsymbol{w},\vec{x})\,\boldsymbol{f}^T(\boldsymbol{w},\vec{x})\,P(\boldsymbol{w},n) \\ &+\eta^2\sum_{i\geq 1}\frac{\lambda_i}{1-\lambda_i}\int d\vec{x}\,d\vec{x}'\,\Phi_i(\vec{x})\,\Psi_i(\vec{x}')\,\Phi_0(\vec{x}')\,\boldsymbol{\nabla}_w^T\boldsymbol{f}(\boldsymbol{w},\vec{x})\boldsymbol{\nabla}_w^T\boldsymbol{f}(\boldsymbol{w},\vec{x}')\,P(\boldsymbol{w},n) + \mathcal{O}(\eta^3)\,. \qquad (4)\end{aligned}$$

To study this evolution equation in the limit $\eta \to 0$, we apply Van Kampen's expansion [12, 7]. We start with the ansatz

$$\boldsymbol{w} = \boldsymbol{\phi}(t) + \sqrt{\eta}\boldsymbol{\xi},$$

with rescaled time $t = \eta n$. This ansatz says that the state of the network $\boldsymbol{w}$ can be described by a deterministic part $\boldsymbol{\phi}(t)$ plus a term of order $\sqrt{\eta}$ containing the fluctuations. The function $\Pi(\boldsymbol{\xi},t) \equiv P(\boldsymbol{\phi}(t) + \sqrt{\eta}\boldsymbol{\xi}, t/\eta)$ is the probability in terms of the new variable $\boldsymbol{\xi}$. From Van Kampen's expansion it immediately follows that the deterministic part $\boldsymbol{\phi}(t)$ has to satisfy the equation

$$\frac{d\boldsymbol{\phi}(t)}{dt} \ = \ \langle \boldsymbol{f}(\boldsymbol{\phi}(t),\vec{x})\rangle_{\vec{x}} \qquad (5)$$

where the average $\langle\ldots\rangle_{\vec{x}}$ is over pattern space. The evolution of $\Pi(\boldsymbol{\xi},t)$ is governed by the Fokker-Planck equation

$$\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial t} \ = \ \mathrm{Tr}\left[H(\boldsymbol{\phi}(t))\boldsymbol{\nabla}_\xi\left[\boldsymbol{\xi}^T\,\Pi(\boldsymbol{\xi},t)\right]\right] \ + \ \frac{1}{2}\mathrm{Tr}\left[\tilde{D}(\boldsymbol{\phi}(t))\boldsymbol{\nabla}_\xi\boldsymbol{\nabla}_\xi^T\,\Pi(\boldsymbol{\xi},t)\right] \qquad (6)$$

with the usual Hessian $H(\boldsymbol{w}) = -\boldsymbol{\nabla}_w\left\langle \boldsymbol{f}^T(\boldsymbol{w},\vec{x})\right\rangle_{\vec{x}}$, butwith the (new) effective diffusion matrix

$$\tilde{D}(\boldsymbol{w}) \ \equiv \ C_0(\boldsymbol{w}) + \lim_{\epsilon\to 1}\sum_{m=1}^{\infty}\left[C_m(\boldsymbol{w}) + C_m^T(\boldsymbol{w})\right]\epsilon^m\,, \qquad (7)$$

where the auto-correlation matrices $C_m$ read

$$\begin{aligned}C_m(\boldsymbol{w}) \ &\equiv \ \sum_{i\geq 1}\lambda_i^m\int d\vec{x}\,d\vec{x}'\,\Phi_i(\vec{x})\,\Psi_i(\vec{x}')\,\Phi_0(\vec{x}')\,\boldsymbol{f}(\boldsymbol{w},\vec{x})\,\boldsymbol{f}^T(\boldsymbol{w},\vec{x}') \\ &= \ \left\langle \boldsymbol{f}(\boldsymbol{w},\vec{x}(m))\,\boldsymbol{f}^T(\boldsymbol{w},\vec{x}(0))\right\rangle_{\vec{x}} - \left\langle \boldsymbol{f}(\boldsymbol{w},\vec{x})\right\rangle_{\vec{x}}\left\langle \boldsymbol{f}^T(\boldsymbol{w},\vec{x})\right\rangle_{\vec{x}}\,. \qquad (8)\end{aligned}$$

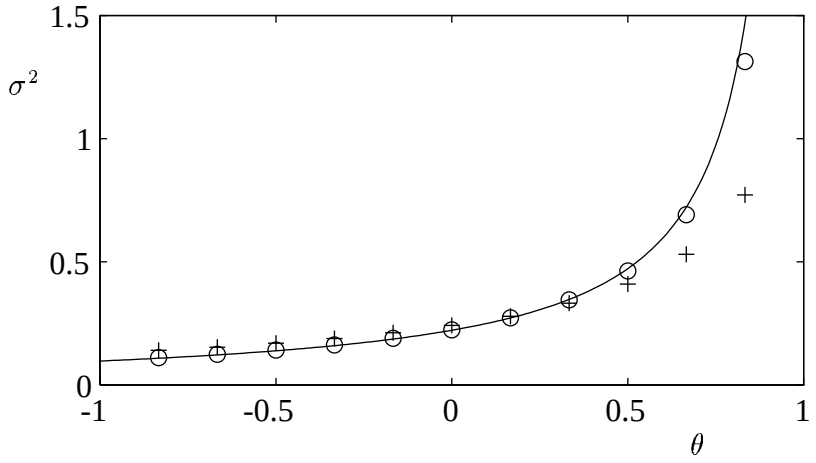


Figure 1: Rescaled asymptotic variance $\sigma^2$ as a function of flip correlation $\theta$. Theoretical prediction (full curve) can be compared with the simulations for $\eta = 0.01$ ("o") and $\eta = 0.1$ ("+"). Standard error bars would be smaller than the size of the symbols.

In deriving (7) and (8) from (4), we used $\lambda_i/(1-\lambda_i) = \lim_{\epsilon\to 1}\sum_{m=1}^{\infty}\lambda_i^m\epsilon^m$, which is correct even if $\lambda_i \neq 1$ lies on the complex unit circle. Equations (7) and (8) constitute our main result. They explain how correlations in the training data affect, through the effective diffusion $\tilde{D}$, the fluctuations in the learning process. For uncorrelated patterns, *i.e.*, if $\rho(\vec{x}|\vec{x}') = \rho(\vec{x})$, all auto-correlation matrices $C_m(\boldsymbol{w}) = 0$ for $m \geq 1$ and the effective diffusion matrix reduces to the usual diffusion matrix $C_0(\boldsymbol{w})$ for on-line learning with random pattern presentation [7]. To generalise our results to higher-order Markov processes, we view a $k$-th order Markov process as a first-order Markov process in the space of extended patterns $\{\vec{x}(m)\} \equiv \{\vec{x}(m), \ldots, \vec{x}(m-k+1)\}$. However, since $\boldsymbol{f}(\boldsymbol{w}, \{\vec{x}(m)\}) = \boldsymbol{f}(\boldsymbol{w}, \vec{x}(m))$, *i.e.*, the weight update does only depend on the last pattern, the auto-correlation matrices in the extended pattern space are equal to the auto-correlation matrices in the original pattern space. Therefore our results are valid for stationary Markov processes of any finite order.

Finally, to illustrate our theory, we simulate the nonlinear Oja learning rule [13]

$$\Delta\boldsymbol{w}(n) \;=\; \eta\,(\vec{x}^T\boldsymbol{w}(n))\,[\vec{x} \;-\; (\vec{x}^T\boldsymbol{w}(n))\,\boldsymbol{w}(n)]$$

in two dimensions, which searches for the principal component of the input correlation matrix $\langle\vec{x}\,\vec{x}^T\rangle_{\vec{x}}$. The absolute value $|x_i|$ is, independent of previous patterns, homogeneously distributed between 0 and $l_i$, with $l_1 = 2$ and $l_2 = 1$. The sign of $x_i$ has a probability $q_i$ to flip after each presentation, *i.e.*, $\vec{x}$ follows a first-order Markov process. The Fokker-Planck equation (6) predicts the asymptotic (rescaled) variance

$$\sigma^2 \;\equiv\; \eta^{-1}\left\langle(\boldsymbol{w} - \langle\boldsymbol{w}\rangle_{w(\infty)})^T(\boldsymbol{w} - \langle\boldsymbol{w}\rangle_{w(\infty)})\right\rangle_{w(\infty)} \;=\; \left\langle\boldsymbol{\xi}^T\boldsymbol{\xi}\right\rangle_{\xi(\infty)} \;=\; \frac{2}{9} + \frac{\theta}{4(1-\theta)}\,.$$

with "flip correlation" $\theta \equiv (1-2q_1)(1-2q_2)$. Simulations are performed with an ensemble of 10 000 independently learning networks, initialised at $\boldsymbol{w}(0) = (1,0)^T$. In the figure it can be seen that the agreement between theory and simulations is better for smaller learning parameters and less correlations, as could be expected. Flip correlation $\theta < 0$ leads to a better sampling of the input space and thus to a smaller asymptotic variance than random pattern presentation ($\theta = 0$).

Our results open new directions for future research. We mention smart sampling techniques to reduce the fluctuations in the learning process and the beneficial influence of time correlations on learning in the presence of local minima [14]. Last but not least, our results may help to understand the learning of (chaotic) time series, which is claimed to be easier if the patterns are presented in their natural order of appearance instead of completely random [15, 16].

***

This work was supported by the Dutch Foundation for Neural Networks and the Japanese Real World Computing Program. We thank one of the referees for a useful suggestion.

## References

[1] D. Rumelhart, J. McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, 1986.

[2] T. Kohonen. *Biol. Cybern.*, 43:59–69, 1982.

[3] H. Ritter and K. Schulten. *Biol. Cybern.*, 60:59–71, 1988.

[4] T. Heskes and B. Kappen. *Phys. Rev. A*, 44:2718–2726, 1991.

[5] G. Radons. *J. Phys. A: Math. Gen.*, 26:3455–3461, 1993.

[6] L. Hansen, R. Pathria, and P. Salamon. *J. Phys. A: Math. Gen.*, 26:63–71, 1993.

[7] T. Heskes. On Fokker-Planck approximations of on-line learning processes. *J. Phys. A: Math. Gen.*, 27:5145–5160, 1994.

[8] N. van Kampen. *Phys. Rep.*, 124:69–160, 1985.

[9] W. Wiegerinck, A. Komoda, and T. Heskes. *J. Phys. A: Math. Gen.*, 27:4425–4437, 1994.

[10] C. Gardiner. *Handbook of Stochastic Methods.* Springer, Berlin, 1985.

[11] S. Gasiorowicz. *Quantum Physics.* Wiley, New York, 1974.

[12] N. van Kampen. *Stochastic Processes in Physics and Chemistry.* North-Holland, Amsterdam, 1992.

[13] E. Oja. *J. Math. Biol.*, 15:267–273, 1982.

[14] T. Heskes, E. Slijpen, and B. Kappen. *Phys. Rev. A*, 46:5221–5231, 1992.

[15] G. Mpitsos and M. Burton. *Neural Networks*, 5:605–625, 1992.

[16] T. Hondou and Y. Sawada. *Prog. Theor. Phys.*, 91:397–402, 1994.